(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENTAMT

(12) **Offenlegungsschrift**

(11) **DE 3931638 A1**

(51) Int. Cl. 5: G10L 5/06

(21) Aktenzeichen: P 39 31 638.6
(22) Anmeldetag: 22. 9. 89
(43) Offenlegungstag: 4. 4. 91

DE 3931638 A1

(71) Anmelder:
Standard Elektrik Lorenz AG, 7000 Stuttgart, DE

(72) Erfinder:
Hackbarth, Heidi, Dr.rer.nat., 7000 Stuttgart, DE;
Immendörfer, Manfred, Dr.rer.nat., 7257 Ditzingen, DE

(54) Verfahren zur sprecheradaptiven Erkennung von Sprache

Ein solches Verfahren soll sowohl zur Erkennung einzelner Wörter als auch von kontinuierlich gesprochener Sprache geeignet sein. Es soll sich durch Robustheit der Wortmustererkennung bei fehlerhafter Silbensegmentierung und bei variabler Aussprache, z. B. bei einem Verschlucken von Silben, auszeichnen. Außerdem muß es eine schnelle Adaption des Systems an einen neuen Sprecher und eine prinzipiell beliebige Generierung und Erweiterung des Wortschatzes aus geschriebenem Text ohne ein explizites Systemtraining durch Vorsprechen ermöglichen. Eine echtzeitnahe Erkennung von Wörtern und Wortfolgen soll auch bei sehr umfangreichen Wortschätzen möglich sein. Bekannte Verfahren zur Spracherkennung benötigen ein sehr aufwendiges Trainingsverfahren. Außerdem wird bei kontinuierlich gesprochener Sprache und größerem Vokabular bereits bei mittleren Vokabulargrößen eine unüberschaubare Hypothesenflut erzeugt.
Erfindungsgemäß wird die Spracherkennung auf der Basis von silbenorientierten Wortuntereinheiten (sogenannten CVC-Einheiten) durchgeführt und es wird ein dreidimensionaler zeitdynamischer Vergleich von Wortmustern aus diesen silbenorientierten Wortuntereinheiten mit Mehrfachhypothesen in einem Testmuster und mit Aussprachevarianten in einem Referenzmuster durchgeführt.

DE 3931638 A1

Beschreibung

Die Erfindung betrifft ein Verfahren zur sprecheradaptiven Erkennung von Sprache. Ein leistungsfähiges Spracherkennungsverfahren hat unter anderem folgende Anforderungen zu erfüllen: Es müssen sowohl isolierte Wörter als auch ein fließender Redetext erkannt werden. Auch bei sehr großen Wortschätzen sollte die Erkennung möglichst in Echtzeit stattfinden. Es ist eine schnelle Adaption an einen neuen Sprecher erforderlich. Eine beliebige Generierung von Referenz-Wörtern und Erweiterung des Wortschatzes soll ohne (gar mehrfaches) Vorsprechen der hinzugefügten Wörter möglich sein. Aussprachevarianten einzelner Wörter müssen automatisch generiert werden können, und zwar ohne explizites Vorsprechen dieser Varianten. Bei fließender Rede soll eine Analyse sich überlappender Worthypothesen die gesprochene Phrase erkennen lassen.

Die bekannten Verfahren zur Spracherkennung aus einem großen Wortschatz (IBM, Dragon, AT&T, BBN, Carëgie-Mellon-Universität (CMU)/Pittsburgh; Übersichtsartikel: Fallside F (1989) Progress in large vocabulary speech recognition. Speech Technology 4(4), 14 – 15) wenden vorwiegend Hidden-Markov-Modelle auf Phonembasis an. In keinem dieser Systeme ist eine automatische Wortschatz-Generierung bzw. -Erweiterung aus geschriebenem Text enthalten. Bei den Erkennern von IBM und Dragon müssen die Wörter isoliert gesprochen werden, während die Erkenner bei AT&T, BBN und CMU nicht sprecheradaptiv arbeiten.

Üblicherweise muß jedes Wort — im Falle einer sprecherabhängigen Erkennung — vom Benutzer ein- oder mehrmals ausgesprochen werden, darüber hinaus — im Fall der sprecherunabhängigen Erkennung — von einer sehr großen Anzahl von Sprechern (Größenordnung 100 bis 1000) mindestens je einmal. Ein solch aufwendiges Trainingsverfahren kann vermieden werden, wenn sprecheradaptive Verfahren verwendet werden. Mit zunehmendem Vokabularumfang ist es hinsichtlich einer echtzeitnahen Spracherkennung notwendig, schnell und ohne großen Rechenaufwand eine kurze Liste wahrscheinlich gesprochener "Wortkandidaten" zu erstellen. Aus diesem Untervokabular aus Wortkandidaten werden anschließend im Zuge der Feinanalyse die gesprochenen Wörter ermittelt. Eine solche Präselektion basiert auf der Klassifikation von groben Merkmalen in Wortuntereinheiten, z.B. in einzelnen Merkmalsvektoren, Phonemen oder Diphonen. Dies stellt für isoliert gesprochene Wörter — auch aus großen Vokabularen — ebenso wie für Ziffernfolgen (vergleiche Chen FR (1986) Lexical access and verification in a broad phonetic approach to continuous digit recognition. IEEE ICASSP, 27.7.1-4; Lagger H, Waibel A (1985) A coarse phonetic knowledge source for template independent large vocabular word recognition. IEEE ICASSP(2), 23.6.1-4; Lubensky D, Feix W (1896) Fast feature-based preclassification of segments in continuous digit recognition. IEEE ICASSP, 27.6.1-4) ein praktikables Verfahren dar. Bei kontinuierlich gesprochener Sprache und größerem Wortschatz führt dies hingegen bereits bei mittleren Vokabulargrößen zu einer unüberschaubaren Hypothesenflut, da prinzipiell bei jeder dieser kleinen Einheiten ein neues Wort anfangen kann und bei jeder Einheit der gesamte Wortvorrat zu durchsuchen wäre. Eine zwei- oder dreidimensionale dynamische Programmierung ist aus Micca G, Pieraccini R, Laface P (1987) Three dimensional DP for phonetic lattice matching. Int Conf on Dig Signal Proc, Firence, Italy und Ruske G, Weigel W (1986) Dynamische Programmierung auf Basis silbenorientierter Einheiten zur automatischen Erkennung gesprochener Sätze. NTG-Fachberichte 94, Sprachkommunikation, 91 – 96 bekannt.

Bei den bekannten Verfahren sind die vorstehend genannten Anforderungen nicht vollständig und teilweise nicht ganz zufriedenstellend erfüllt.

Der Erfindung liegt die Aufgabe zugrunde, ein Verfahren zur sprecheradaptiven Erkennung von Sprache zu schaffen, das in der Lage ist, sowohl isolierte Wörter als auch kontinuierliche Sprache bei einem praktisch unbegrenzten Vokabular echtzeitnah zu erkennen und das auch die weiteren Anforderungen an ein leistungsfähiges Spracherkennungsverfahren erfüllt.

Diese Aufgabe wird erfindungsgemäß durch das Verfahren nach Patentanspruch 1 gelöst.

Weiterbildungen der Erfindung sind den Unteransprüchen zu entnehmen. Die Vorteile der Erfindung liegen insbesondere in der Robustheit der Wortmustererkennung bei fehlerhafter Silbensegmentierung und bei variabler Aussprache, z.B. beim Verschlucken von Silben. Große Referenzwortschätze brauchen nicht explizit vorgesprochen zu werden. Silbenorientierte Wortuntereinheiten ermöglichen im Vergleich zu den sonst gebräuchlichen Phonemen eine effizientere Generierung von Worthypothesen.

Ein Ausführungsbeispiel der Erfindung wird im folgenden anhand der Zeichnung erläutert. Es zeigen:

Fig. 1 ein Funktionsdiagramm, das den modularen Aufbau des erfindungsgemäßen Verfahrens erkennen läßt,

Fig. 2 ein Diagramm zur Erläuterung des dreidimensionalen zeitdynamischen Vergleichs zur Worterkennung, und

Fig. 3 ein Funktionsdiagramm zur Erläuterung der akustischen Präselektion eines Untervokabulars bei der Erkennung isolierter Wörter oder fließender Rede.

Bei einem zur erkennenden Sprachsignal 1 findet zunächst eine Extraktion 2 von Merkmalsvektoren statt. Ein solcher Merkmalsvektor wird z.B. aus Filterbank-Koeffizienten gebildet, die die Intensitäten für die verschiedenen Frequenzbereiche des Signals kennzeichnen. Anschließend wird eine automatische Segmentierung und Klassifikation 3 der aufeinanderfolgenden Merkmalsvektoren durchgeführt, und zwar in silbenorientierte Wortuntereinheiten. Geeignete silbenorientierte Wortuntereinheiten sind z.B. CVC-Einheiten (CVC für: consonant cluster — vocalic syllable kernel — consonant cluster), die aus einem vokalischen Silbenkern V mit vorausgehender silbeninitialer und nachfolgender silbenfinaler Konsonantenfolge oder einzelnen konsonantischen Phonemen C je Silbe bestehen. Die Segmentierung und Klassifikation 3 der Vektorfolgen wird anhand eines gespeicherten Vorrats an Wortuntereinheiten, im folgenden als Wortuntereinheiten-Inventar 4 bezeichnet, durchgeführt. Die Segmentierung und Klassifikation 3 der Vektorfolgen ergibt ein Hypothesennetz 6 (oder auch Netzwerk) aus Wortuntereinheiten, das einer Worterkennung 7 zugeführt wird.

Ein Wortschatz 8 enthält abgespeicherte Referenzmuster von Wörtern. In dem Verfahrensschritt Worterkennung 7 wird aus dem Hypothesennetz 6 aus Wortuntereinheiten unter Zugriff auf die abgespeicherten Referenzmuster ein Netz 10 von Worthypothesen regeneriert. Diese Worthypothesen werden sich bei kontinuierlicher Sprache im allgemeinen überlappen; aus ihnen wird in einem nachfolgenden Syntax-Schritt 12 die gesprochene Phrase oder der gesprochene Satz ermittelt.

In einem Verfahrenschritt Sprecheradaptivität 13 wird in einer kurzen Trainingsphase das Spracherkennungsverfahren an einen neuen Benutzer angepaßt, ohne daß dieser den gesamten Wortschatz vorsprechen muß. Dieser Verfahrensschritt wird als Hybridansatz durchgeführt, d.h. er wird sowohl auf die Ebene der Merkmalsvektoren als auch auf die Ebene der Wortuntereinheiten angewendet.

Der in dem Verfahren verwendete Wortschatz 8 wird durch die Eingabe von geschriebenem Text 14 erstellt und erweitert. Die Grapheme dieses Textes werden in einer Graphem-Umsetzung 15 automatisch in die hier verwendete Wortuntereinheiten-Notierung der Wörter umgewandelt. Die gleichfalls erzeugten Aussprachevarianten werden ebenfalls in diese Wortuntereinheiten-Notierung umgesetzt.

Um das Suchen in großen Wortschätzen zu beschleunigen, ist eine Präselektion 16 vorgesehen, mit deren Hilfe lediglich ein ausgewähltes Untervokabular auf Ähnlichkeit mit der gesprochenen Äußerung untersucht wird.

Die Verfahrensschritte oder Module Worterkennung 7 und Wortschatz 8 werden nun anhand von Fig. 2 eingehender erläutert. Die Worterkennung 7 wird durchgeführt, indem das Hypothesennetz 6 aus Wortuntereinheiten des Testmusters mit den Referenzmustern im Wortschatz 8 verglichen werden. In diesen Referenzmustern oder Wortmodellen sind neben der Standardaussprache des jeweiligen Wortes auch Aussprachevarianten, und zwar Lineare Varianten einzelner Wortuntereinheiten oder Varianten mit Silbenauslassungen, integriert. In dem Wortschatz 8 (Fig. 2) ist dies beispielshaft anhand des Wortes "Erdbeeren" dargestellt: Die Standardaussprache V1 als dreisilbiges Wort, eine (lineare) Variante V2 an einer Stelle, sowie eine Silbensprung-Variante V3.

Sowohl als Referenzmuster aus dem Wortschatz 8 wie auch als Testmuster liegt je ein Wortuntereinheiten-Netz vor. Zur Worterkennung muß deshalb ein dreidimensionaler zeitdynamischer Vergleich 18 durchgeführt werden, bei dem zwei Dimensionen durch die zeitliche Entwicklung von Test- und Referenzmuster gegeben sind, während die dritte Dimension von den verschiedenen Hypothesen oder Aussprachevarianten pro Wortuntereinheit aufgespannt wird.

Es sind zwar schon Spracherkennungs-Verfahren mit dreidimensionalem Vergleich bekannt, sie verarbeiten aber höchstens zwei Alternativen pro Wortuntereinheit und beruhen insbesondere auf einer Segmentierung der Sprachsignale in Folgen von Phonemen. Dies hat eine ganz erhebliche Anzahl von möglichen Zuordnungen zur Folge. Die in dem erfindungsgemäßen Verfahren verwendeten silbenorientierten Wortuntereinheiten bieten dagegen den Vorteil, daß bei der zeitdynamischen Musteranpassung nur Einfügungen oder Auslassungen von ganzen Silben vorkommen können, z.B. von einem Vokal zur silbenfinalen Konsonantenfolge der nachfolgende Silbe (aus CVC/CVC wird CVC). Dies hat eine erhebliche Einschränkung der möglichen Zuordnungen im Vergeich zu den bekannten Verfahren zur Folge.

Um das Vokabular aus einem geschriebenen Text automatisch zu erstellen und zu erweitern, wird die Orthografie — auch Rechtschrift oder Graphemfolge eines neuen Wortes umgewandelt in eine Folge von Indizes von silbenorientierten Wortuntereinheiten. Diese entsprechen den Indizes der Elemente des Inventars 4, das in der Worterkennung 7 als Referenz zur Klassifikation der akustischen oder gesprochenen Wortuntereinheiten verwendet wird. Die Referenz-Wortuntereinheiten werden in der Trainingsphase aus markierten Sprachdaten gewonnen, die alle vorkommenden Wortuntereinheiten enthalten. Ein Worteintrag in das Vokabular enthält demgemäß neben der Orthografie, Silbenzahl usw. solche Indexfolgen für die Standardaussprache und die Aussprachevarianten. Während der Worterkennung werden diese Indexfolgen mit dem Hypothesennetz aus Wortuntereinheiten — die ebenfalls in Indexform vorliegen — verglichen (Fig. 2). Entscheidend ist hier die Kompatibilität zwischen der Verarbeitung des Sprachsignals zu Wortuntereinheiten und der damit übereinstimmenden Analyse des geschriebenen Textes.

Um sowohl der hohen Variabilität der Aussprache eines einzelnen Benutzers und erst recht der Aussprache verschiedener Benutzer Rechnung zu tragen, ist es im Hinblick auf eine zuverlässige Spracherkennung außerdem vorteilhaft, Aussprachevarianten zu berücksichtigen. Bei umfangreichen Wortschätzen ist nur die automatische Generierung solcher Aussprachevarianten mit Hilfe phonologischer Regeln praktikabel.

Um den Suchvorgang in umfangreichen Wortschätzen zu beschleunigen, wird eine Präselektion 18 (Fig. 3) angewendet, mit deren Hilfe lediglich ein ausgewähltes Untervokabular auf Ähnlichkeit mit der gesprochenen Äußerung untersucht wird. Die Präselektion beruht auf einer Klassifikation 19 nach "groben" silbenorientierten Wortuntereinheiten und einer "groben" und robusten Suche (Worterkennung) 20 in einem Wortschatz 21, der entsprechend "grob" kodierte Einträge enthält. Das Referenzmaterial zu der Identifikation der groben Wortuntereinheiten, ein sogenanntes Grob-Inventar 22, wird aus dem alle Wortuntereinheiten umfassenden Inventar 4 durch Klassenbildung generiert, die getrennt je nach Typ der Wortuntereinheit erfolgt, z.B. jeweils Vokale, silbeninitiale oder -finale Konsonatenfolgen.

Dabei werden akustisch ähnliche Wortuntereinheiten in sogenannten Clustern zusammengefaßt. Dies kann z.B. wahlweise durch eine akustische Beurteilung, durch eine Ermittlung disjunkter Untermengen auf der Basis von Ähnlichkeits- oder Verwechslungsmatrizen und/oder mit Hilfe bekannter Clusteringverfahren erfolgen.

Das Zwischenergebnis nach der Grob-Klassifikation entsprechend dem groben Referenzinventar besteht also aus einer Folge 24 aus groben Wortuntereinheiten. Aus dieser Folge 24 ermittelt das Modul zur groben Worterkennung 20 ein Untervokabular 25 mit den bestpassenden Wortkandidaten. Deren Wortmodelle, d.h. deren Wortuntereinheiten-Notierungen für die Standardaussprache und die Aussprachevarianten werden in der Worterkennung 7 zum Vergleich mit dem Hypothesennetz 6 herangezogen und nochmals eine Auswahl 16 getroffen.

Die beschriebene Präselektion eignet sich im Gegensatz zu allen bisher bekannten Methoden sowohl zur schnellen Vorauswahl eines Untervokabulars bei der Erkennung von Einzelwörtern als auch von verbundener Sprache, da die Generierung von Wortkandidaten auf die Silbenanfänge reduziert ist und somit eine überschaubare Hypothesenmenge erzeugt.

Patentansprüche

1. Verfahren zur sprecheradaptiven Erkennung von Sprache, **dadurch gekennzeichnet,**

— daß aus dem zu erkennenden Sprachsignal Merkmalsvektoren extrahiert werden,

– daß die aufeinanderfolgenden extrahierten Merkmalsvektoren in silbenorientierte Wortuntereinheiten segmentiert und klassifiziert werden, und

– daß mit diesen Wortuntereinheiten mit Mehrfachhypothesen aus einem abgespeicherten Testmuster-Inventar ein dreidimensionaler zeitdynamischer Vergleich mit Aussprachevarianten aus einem Referenzmuster-Wortschatz durchgeführt wird.

2. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß bei fließender Rede die sich überlappenden Worthypothesen einer syntaktischen Analyse unterworfen und dadurch die gesprochene Phrase ermittelt wird.

3. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß die abgespeicherten Referenz-Sprachdaten mit einem Hybridansatz anhand der in einer kurzen Trainingsphase gesprochenen Äußerungen eines neuen Sprechers an diesen Sprecher adaptiert werden.

4. Verfahren nach Anspruch 3, dadurch gekennzeichnet, daß sowohl die Merkmalsvektoren als auch die Wortuntereinheiten adaptiert werden.

5. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß der abgespeicherte Wortschatz einschließlich Aussprachevarianten durch Eingeben von geschriebenem Text und regelbasiertes Umsetzen dieses Textes in Symbole für Wortuntereinheiten generiert und erweitert wird.

6. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß zum beschleunigten Erkennen von Sprache bei großen gespeicherten Wortschätzen eine Präselektion eines Untervokabulars mit Hilfe von silbenorientierten Wortuntereinheiten durchgeführt wird.

Hierzu 3 Seite(n) Zeichnungen

Nummer: DE 39 31 638 A1
Int. Cl.[5]: G 10 L 5/06
Offenlegungstag: 4. April 1991

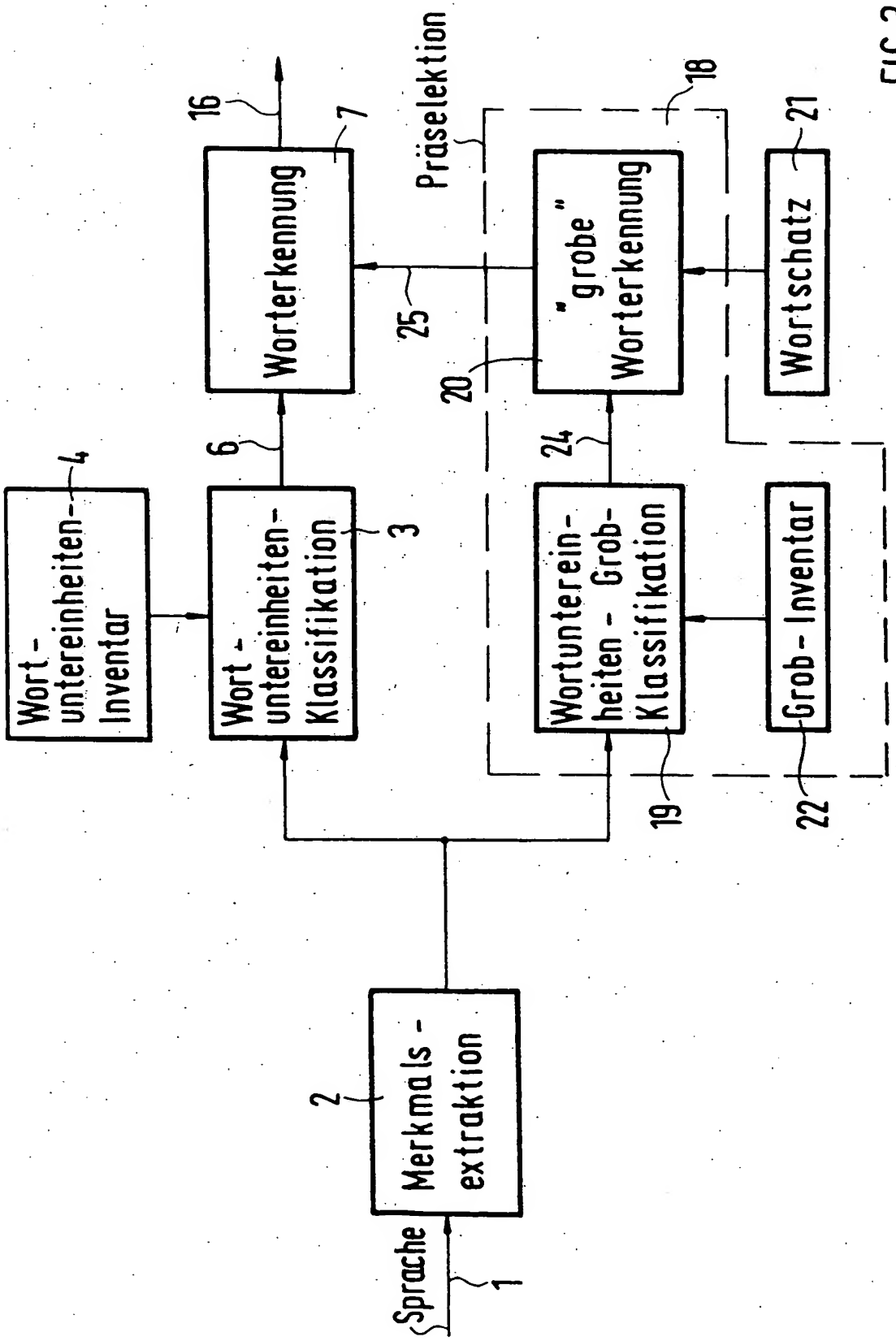

FIG.3

Nummer: DE 39 31 638 A1
Int. Cl.[5]: G 10 L 5/06
Offenlegungstag: 4. April 1991

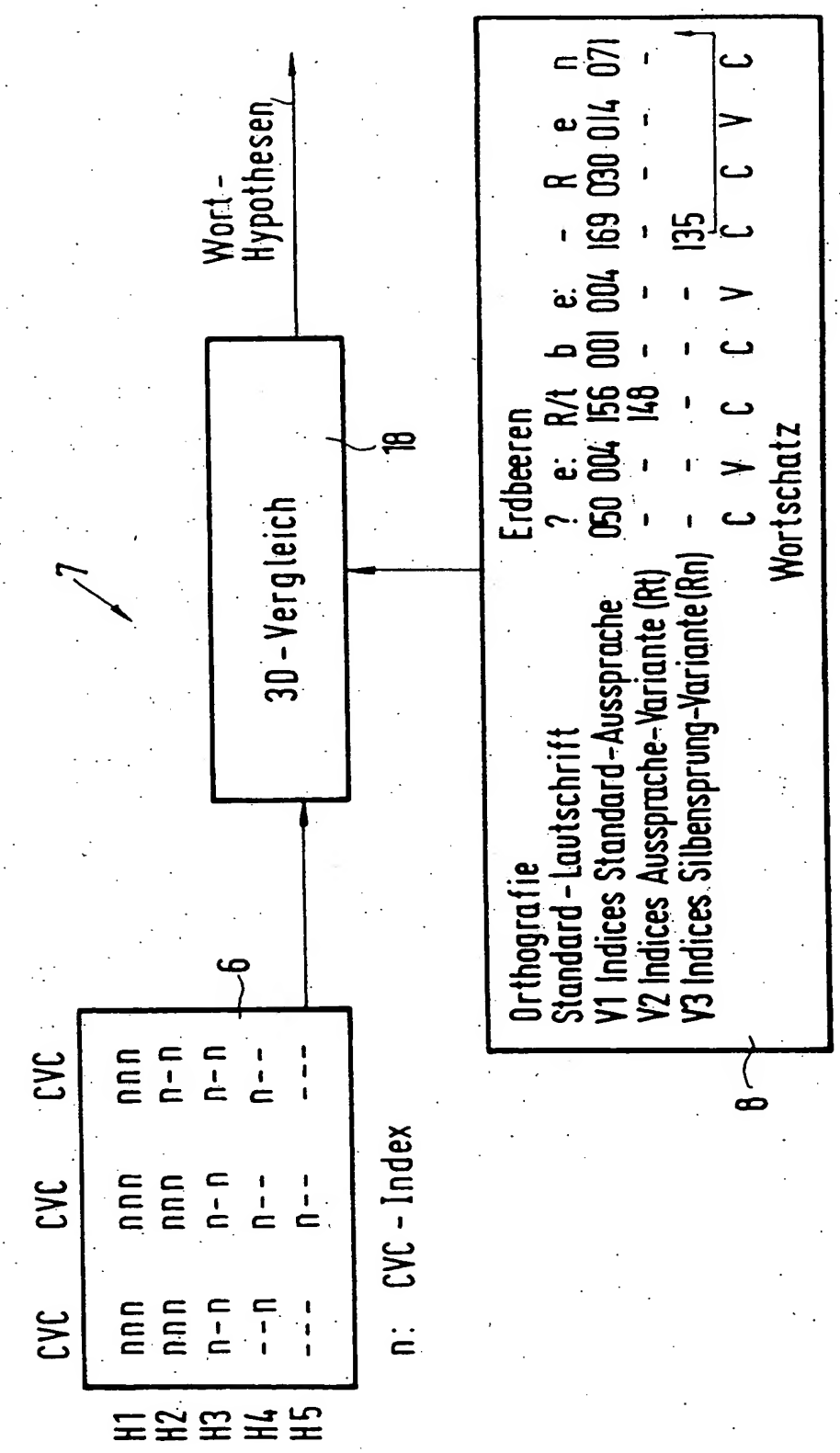

FIG.2

Nummer: DE 39 31 638 A1
Int. Cl.[5]: G 10 L 5/06
Offenlegungstag: 4. April 1991

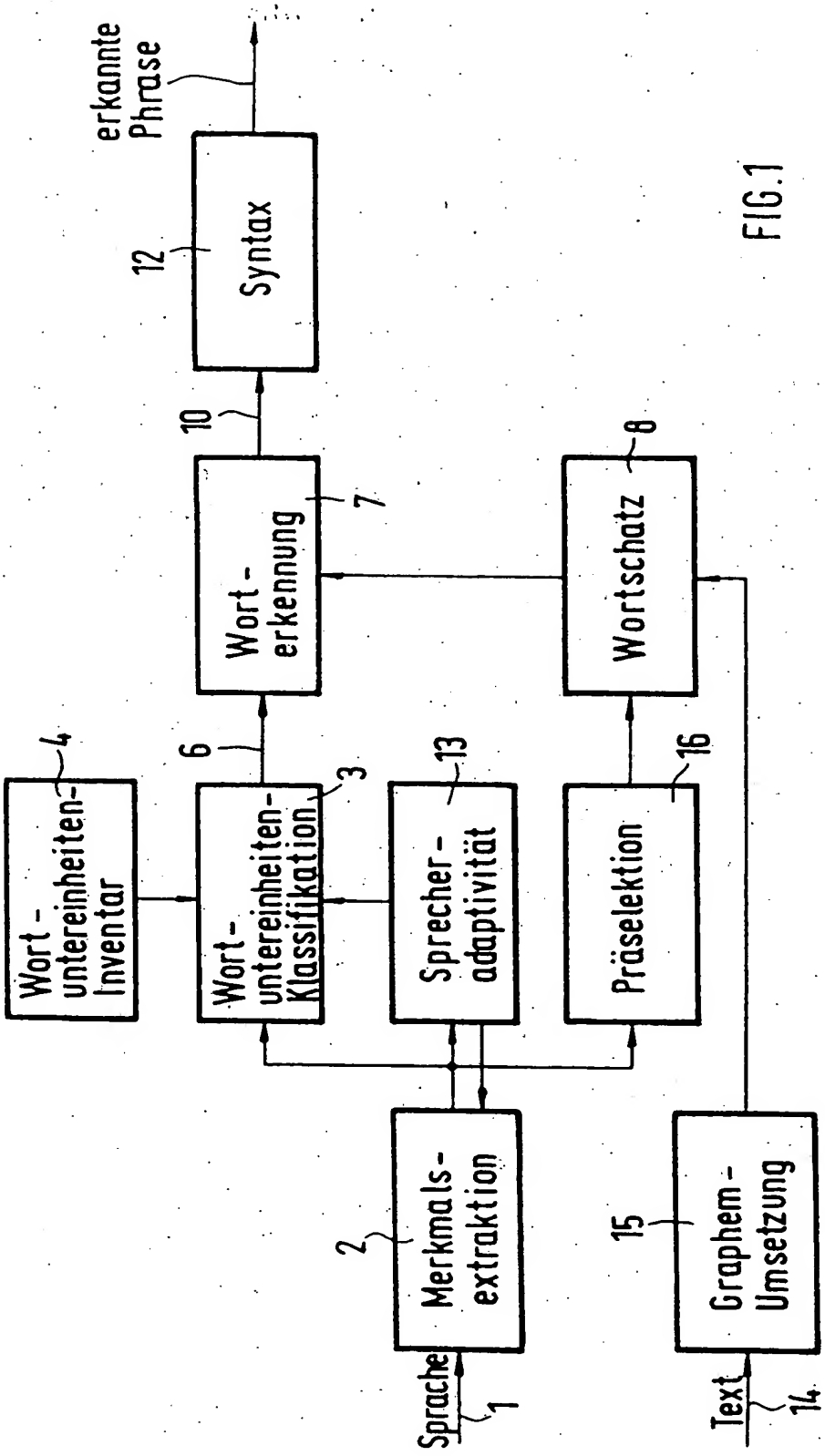

FIG.1

— Leerseite —